Utilizing shRNA screening data

This document describes the data generated from screening human cancer cell lines for proliferation effects with a pooled shRNA library of 54,020 shRNAs, and how to replicate or extend results first reported in Cheung, H.W. *et al.*, (2011). *Proc. Natl. Acad. Sci USA*, **108**(30):12372-7.

Data Generation

Lentiviral pLKO.1- shRNA constructs were obtained from the RNAi Consortium, and the human 54K pool of 54,020 shRNA plasmids was assembled by combining 16 normalized subpools of ˜3400 shRNA plasmids. The list of 54,020 shRNAs can be found at http://www.broadinstitute.org/igp. Genome-scale pooled shRNA screens to identify genes essential for proliferation in human cancer cell lines were performed using a lentivirally delivered pool of 54,020 shRNAs targeting 11,194 genes. The culture conditions for all cancer cell lines are listed in Cheung, H.W. *et al.*, Table S1. Each cell line was infected in quadruplicate and propagated for at least 16 population doublings. The abundance of shRNA constructs relative to the initial DNA plasmid pool was measured by microarray hybridization and analyzed by using a uniform pipeline. Detailed descriptions of each procedure can be found in Cheung, H.W. *et al.*, SI Methods.

Data Processing

Raw .CEL files from custom Affymetrix barcode arrays were processed with a modified version of dCHIP software (Luo *et al*, (2008). *Proc Natl Acad Sci U S A.* **105**(51):20380-5). GenePattern modules (available at the Broad Institute Integrative Genomics Portal http://www.broadinstitute.org/igp) were used to calculate the $\log_2$ fold change in shRNA abundances for each cell line at the conclusion of the screening relative to the initial plasmid DNA reference pool (available as PLASMID_DNA.geo.txt) and to normalize these depletion values by using peak median absolute deviation normalization, a variation of Z score with median absolute deviation. The values reported in each cell-specific file represent dCHIP and peak median absolution deviation normalized depletion values for each shRNA per biological replicate.

How to Use these Data

To replicate or extend the results reported in Cheung, H.W. *et al.*, you will need to first compute log fold change values for each cell line. To do so, use the following steps:

1. For each shRNA, obtain the mean of the replicate measurements in the file PLASMID_DNA.geo.txt. These measurements are on the plasmid DNA of the pooled shRNAs and represent the starting concentration of each shRNA.

2. For each shRNA, obtain the mean of the replicate measurements in a specific cell line from the associated cell line specific data file (i.e. files named CELL_LINE.geo.txt). These measurements represent the final time point in the screening experiment, when shRNAs are likely to have been depleted or enriched as a result of shRNA effects on cellular proliferation.

3. To compute $\log_2$ fold change values, subtract the final time point values (item 2 above) from the starting concentration (item 1 above). This is the $\log_2$ fold change value for each shRNA.

.

After computing $\log_2$ fold change values, you can compute gene level scores and rank order genes using our analysis and visualiztion tool GENE-E (http://www.broadinstitute.org/cancer/software/GENE-E). This program includes RIGER (RNAi gene enrichment ranking, Luo *et al.*, (2008)), three complementary methods for collapsing shRNA scores to gene scores. These methods include (i) ranking genes by their highest shRNA depletion score, (ii) ranking genes based on the P value rank (correcting for different set sizes of shRNA targeting different genes) of their second best ranked shRNA, and (iii) ranking genes using a KS statistic in an approach similar to gene set enrichment analysis. Detailed descriptions of each procedure can be found in Cheung, H.W. *et al.*, SI Methods.

Online Resources

Access to additional data, analysis tools and tutorials for data use can be found here:

http://www.broadinstitute.org/IGP/home